CS 224S / Linguist 285
Spoken Language Processing

Andrew Maas | Stanford University | Spring 2025

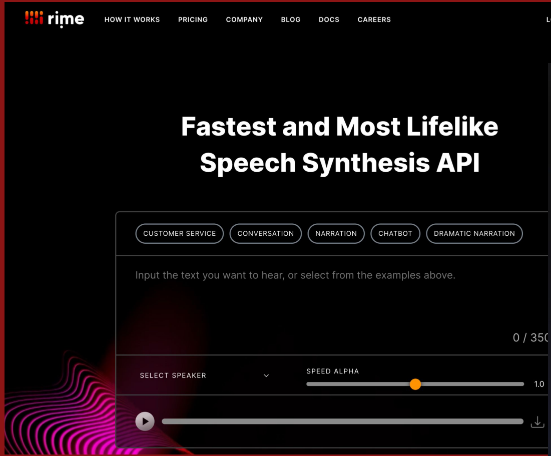
Lecture 3: Text-to-Speech (TTS) Overview

Outline

- Modern speech synthesis (= TTS). State of the art + challenges
- Prosody and Intonation
- Early speech synthesis systems
- TTS Overview
  - The basic modules
  - Text Analysis

  - (Next class) Waveform synthesis

Modern TTS systems are quite good!



rime HOW IT WORKS PRICING COMPANY BLOG DOCS CAREERS LOG IN

Fastest and Most Lifelike Speech Synthesis API

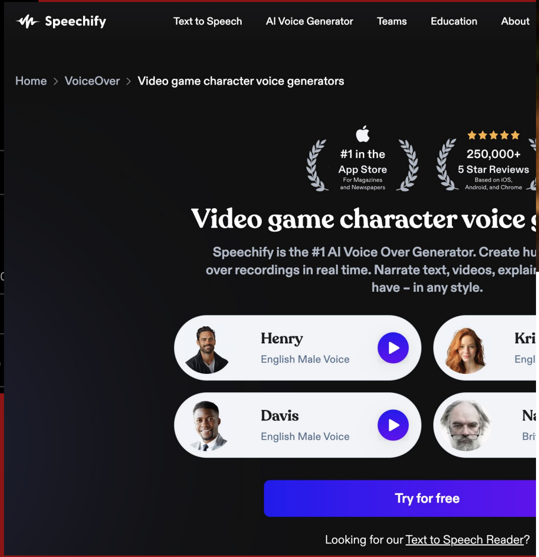
CUSTOMER SERVICE CONVERSATION NARRATION CHATBOT DRAMATIC NARRATION

Input the text you want to hear, or select from the examples above.

0 / 350

SELECT SPEAKER

SPEED ALPHA 1.0



Speechify Text to Speech AI Voice Generator Teams Education About

Home > VoiceOver > Video game character voice generators

#1 in the App Store For Magazines and Newspapers

★★★★★ 250,000+ 5 Star Reviews Based on iOS, Android, and Chrome

Video game character voice g

Speechify is the #1 AI Voice Over Generator. Create hu over recordings in real time. Narrate text, videos, explai have – in any style.

Henry English Male Voice

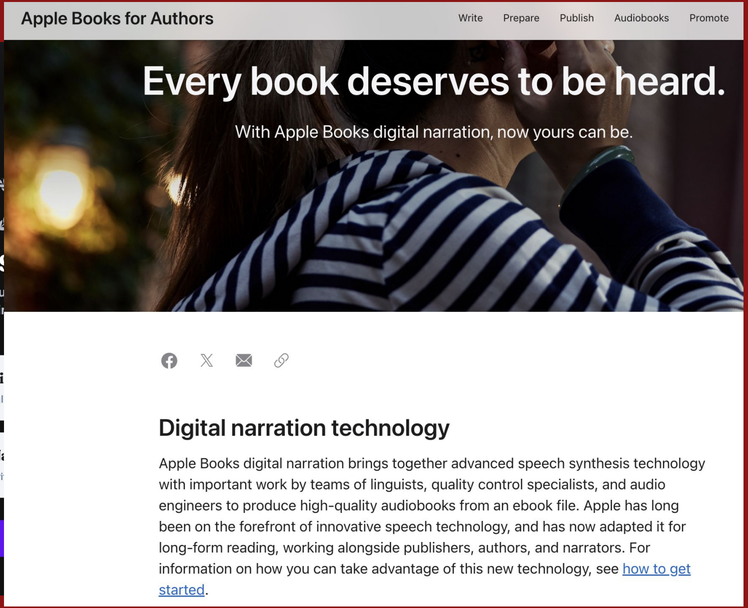
Kri Engl

Davis English Male Voice

Na Bri

Try for free

Looking for our Text to Speech Reader?



Apple Books for Authors Write Prepare Publish Audiobooks Promote

Every book deserves to be heard.

With Apple Books digital narration, now yours can be.

Digital narration technology

Apple Books digital narration brings together advanced speech synthesis technology with important work by teams of linguists, quality control specialists, and audio engineers to produce high-quality audiobooks from an ebook file. Apple has long been on the forefront of innovative speech technology, and has now adapted it for long-form reading, working alongside publishers, authors, and narrators. For information on how you can take advantage of this new technology, see how to get started.

Rime.ai

Speechify.com

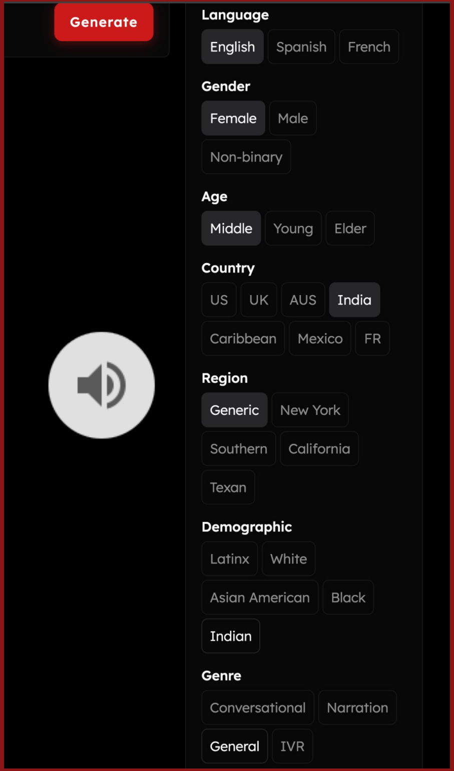
https://authors.apple.com/support/4519-digital-narration-audiobooks

Many cloud provider APIs and public models

Recent product innovation in TTS: Rime.ai

Deep learning models with natural voices allow for a much larger range of speakers and expressive style in training data. Many ways to productize by choosing applications and building custom models.

Rime.ai leverages deep learning TTS to provide many styles of voice to match regions and demographics



Language: [✓ English] Spanish French

Gender: [✓ Female] Male Non-binary

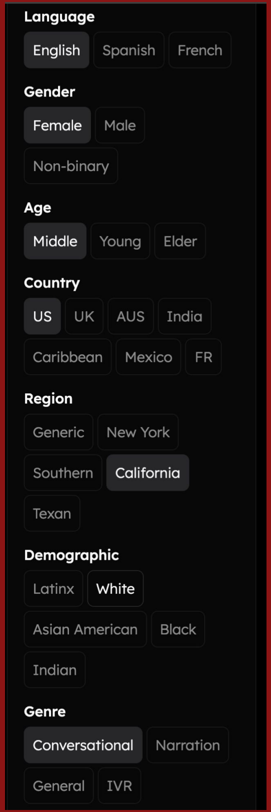
Age: [✓ Middle] Young Elder

Country: US UK AUS [✓ India] Caribbean Mexico FR

Region: [✓ Generic] New York Southern California Texan

Demographic: Latinx White Asian American Black [✓ Indian]

Genre: Conversational Narration [✓ General] IVR



Language: [✓ English] Spanish French

Gender: [✓ Female] Male Non-binary

Age: [✓ Middle] Young Elder

Country: [✓ US] UK AUS India Caribbean Mexico FR

Region: Generic New York Southern [✓ California] Texan

Demographic: Latinx [✓ White] Asian American Black Indian

Genre: [✓ Conversational] Narration General IVR

Where can modern systems improve?

- Emotional expression
- Inferring expression from text
- Singing, accents, and controllability
- Realtime factor and low-latency for spoken dialog systems
- Hardware constraints, cloud vs on-device synthesis
- The dream:
    - Large range of voices and emotional expression
    - Controllable voice types
    - Works on any text. Allows for direct control of expression OR inferring from text / context
    - Runs on smartphone without internet

What does a partly-working TTS system sound like?

Sample 1

Sample 2

Sample 3

Sample 4

Prosody and Intonation

Defining Intonation

- **Ladd (1996) “Intonational phonology”**

  The use of suprasegmental phonetic features to convey sentence-level pragmatic meanings

- **“The use of suprasegmental phonetic features [...]**
  - Suprasegmental = above & beyond the segment/phone
    - F0 (pitch)
    - Intensity (energy)
    - Duration

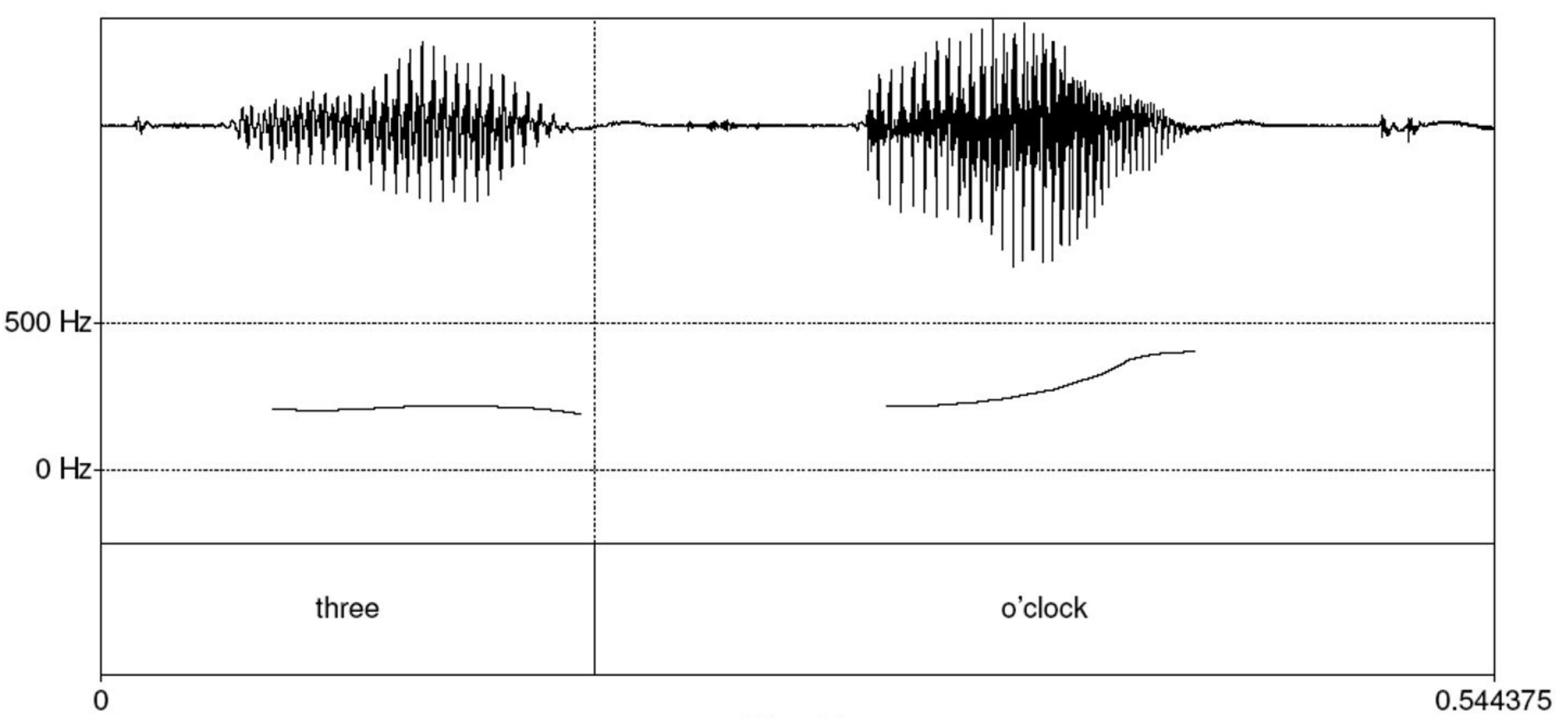
- **to convey sentence-level pragmatic meanings”**
  - i.e. meanings that apply to phrases or utterances as a whole, not lexical stress, not lexical tone.

Prosody and Intonation in TTS: What must we specify?

- **Prominence/Accent:** decide which words are accented, which syllable has accent, what sort of accent
- **Boundaries:** Decide where intonational boundaries are
- **Duration:** Specify length of each segment
- **F0:** Generate F0 contour from these

Even joint deep learning TTS systems do this, but in some cases they are implicitly setting these parameters without direct controllability / parameterization

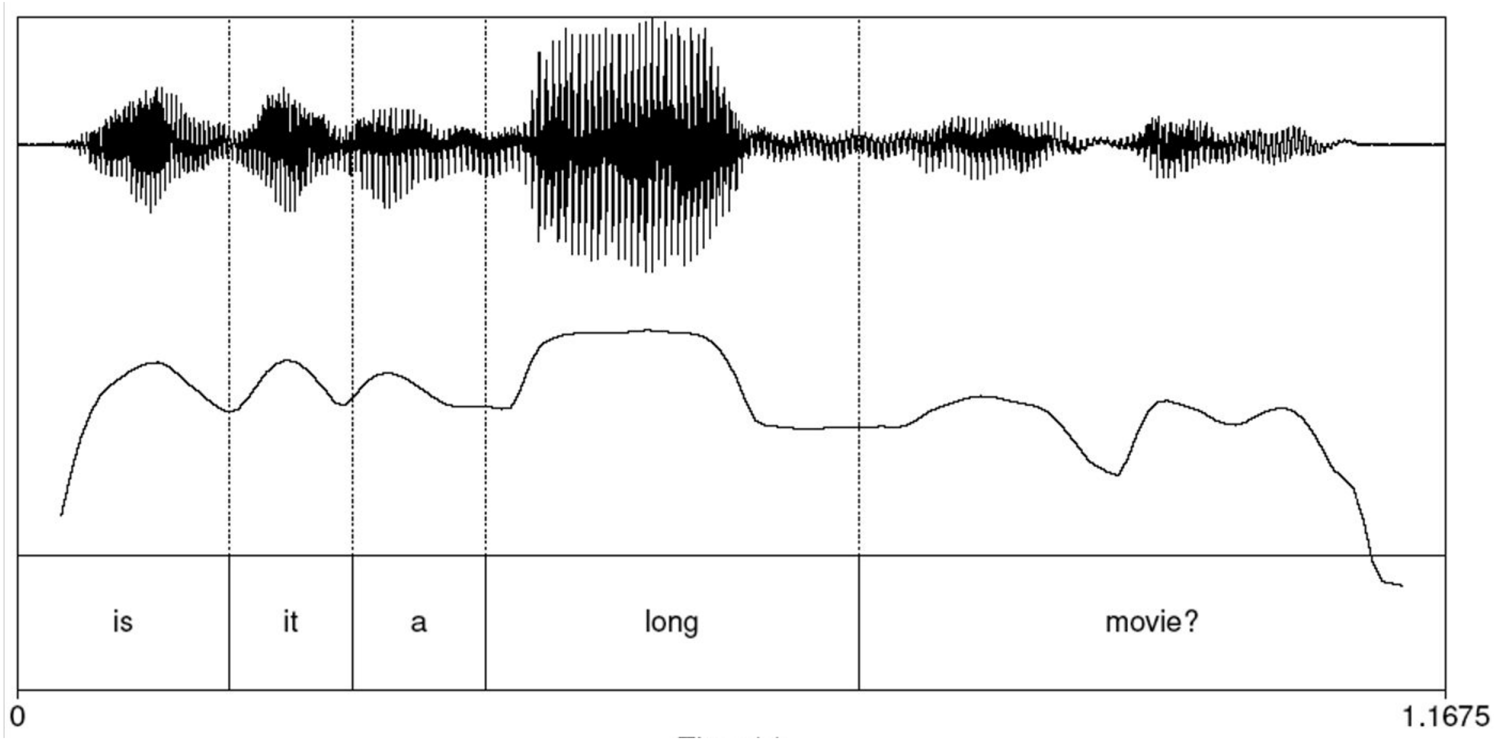
Pitch Track: Conveying meaning with pitch



Pitch is not frequency

- **Pitch is the mental sensation or perceptual correlate of F0**
- **Relationship between pitch and F0 is not linear;**
  - human pitch perception is most accurate between 100Hz and 1000Hz.
    - Linear in this range
    - Logarithmic above 1000Hz
- **Mel scale is one model of this F0-pitch mapping**
  - A mel is a unit of pitch defined so that pairs of sounds which are perceptually equidistant in pitch are separated by an equal number of mels
  - Frequency in mels = 1127 ln (1 + f/700)

Pitch and accent: What might you infer from the prosody?



Prosodic boundaries

I met Mary and Elena’s mother at the mall yesterday.

🔊

I met Mary, and Elena’s mother at the mall yesterday.

🔊

French [bread and cheese]

🔊

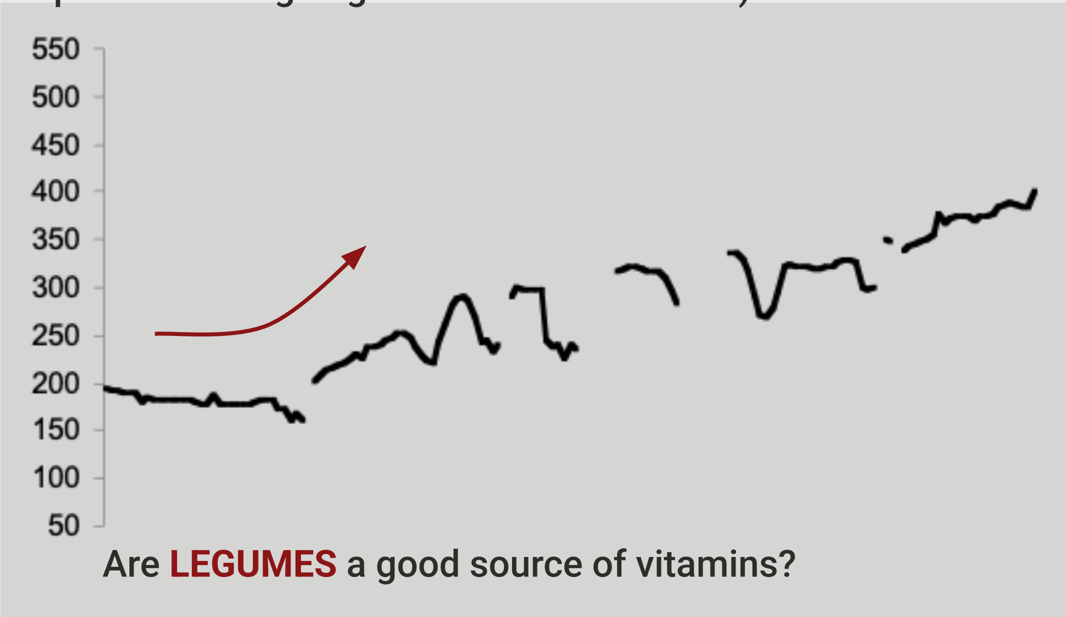
[French bread] and [cheese]

🔊

Yes-No Question Tune

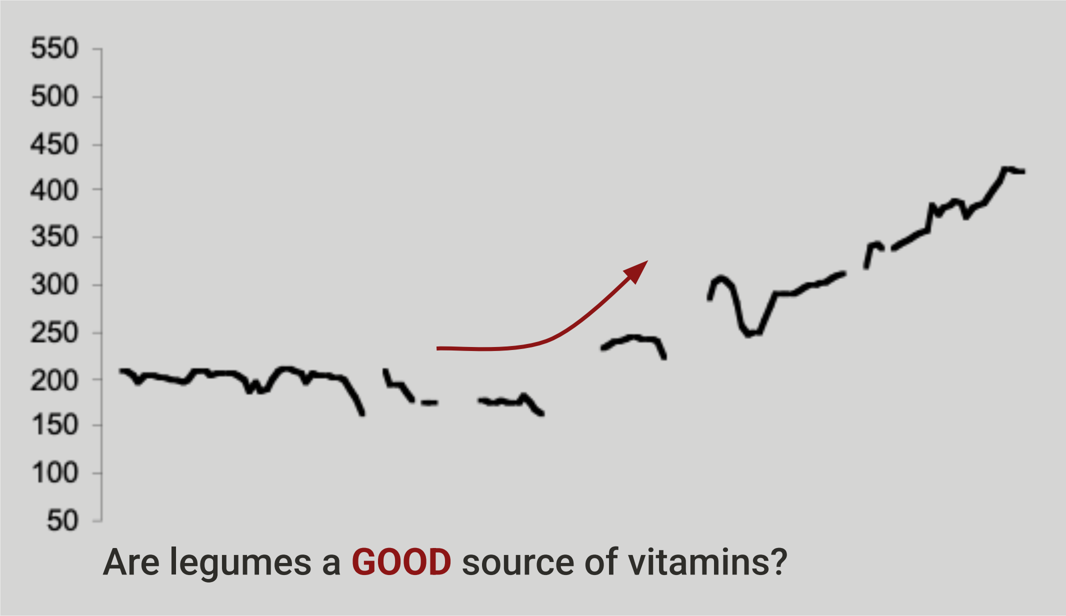
- **Rise** from the main accent to the end of the sentence

  (This is a pitch track. A plot of the "signing tune" of the utterance)



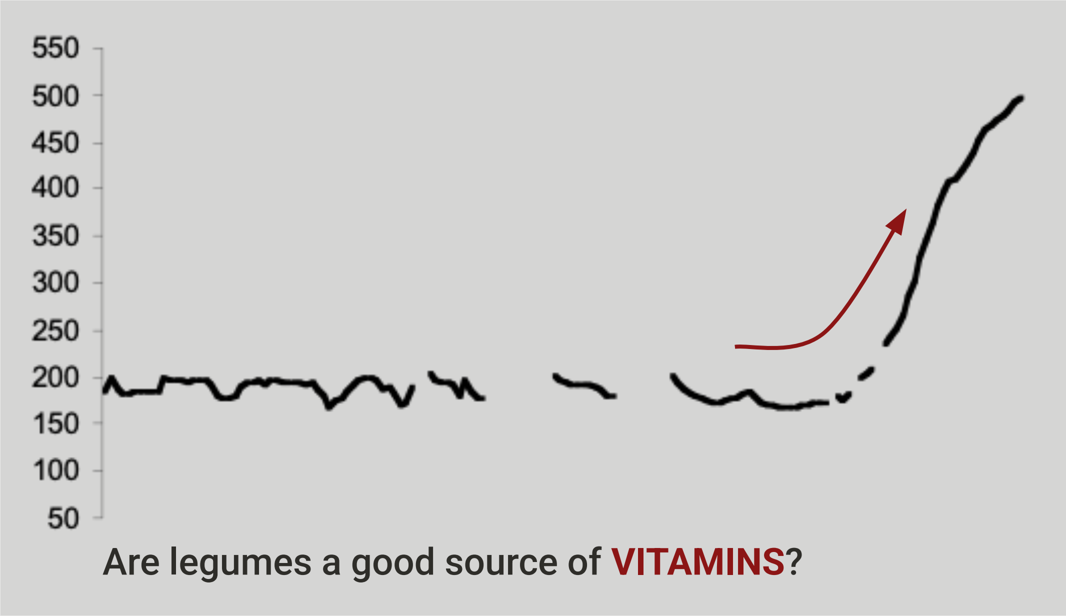
Yes-No Question Tune

- **Rise** from the main accent to the end of the sentence



Yes-No Question Tune

- **Rise** from the main accent to the end of the sentence



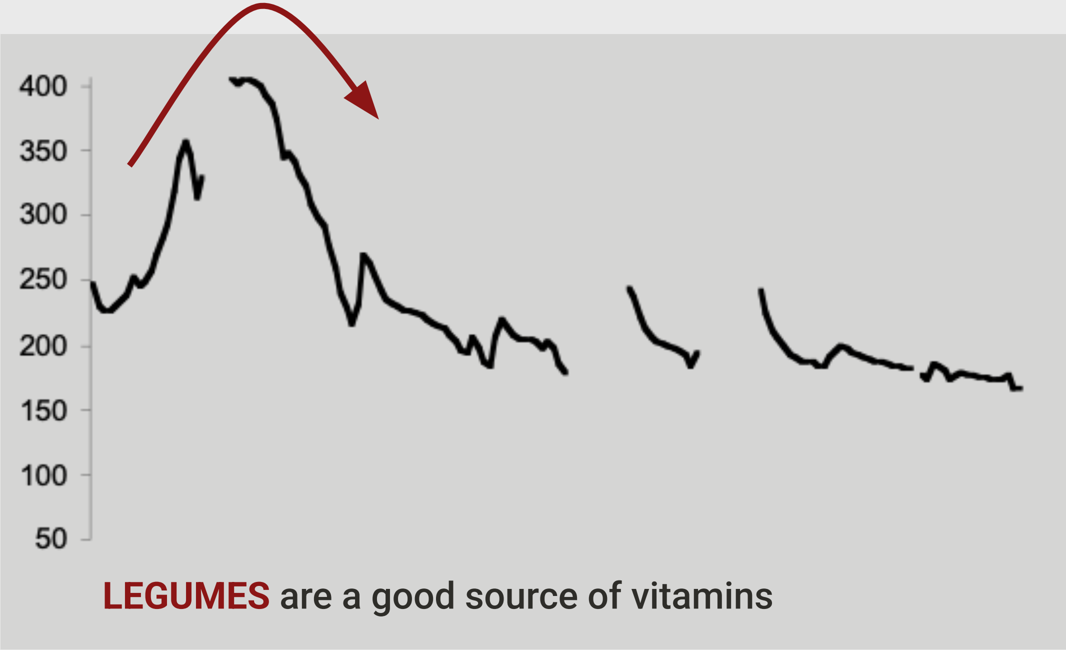
Stress vs. Accent

- **Stress:** structural property of a word
    - ○ Fixed in the lexicon: marks a potential (arbitrary) location for an accent to occur, if there is one
- **Accent:** property of a word in context
    - ○ Context-dependent. Marks important words in the discourse

(x)					(x)		(accented syll)
x					x		stressed syll
x			x		x		full vowels
x	x	x	x	x	x	x	syllables
vi	ta	mins	Ca	li	for	nia	

Same ‘Tune’, Different Alignment

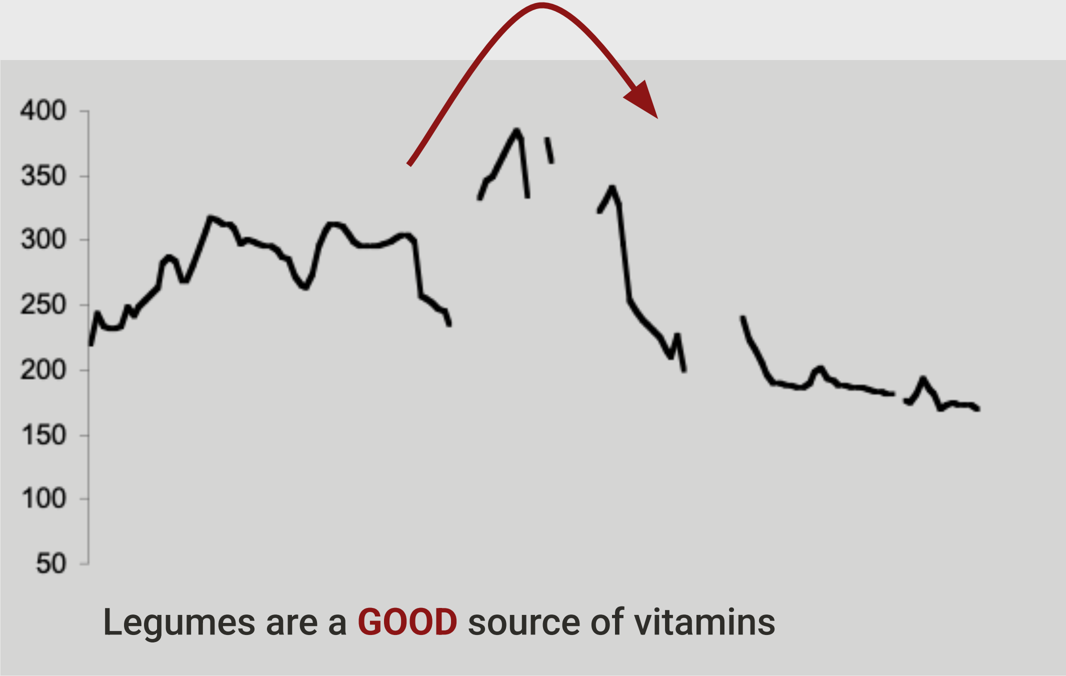
- The main **rise-fall** accent (= “I assert this”) shifts locations



LEGUMES are a good source of vitamins

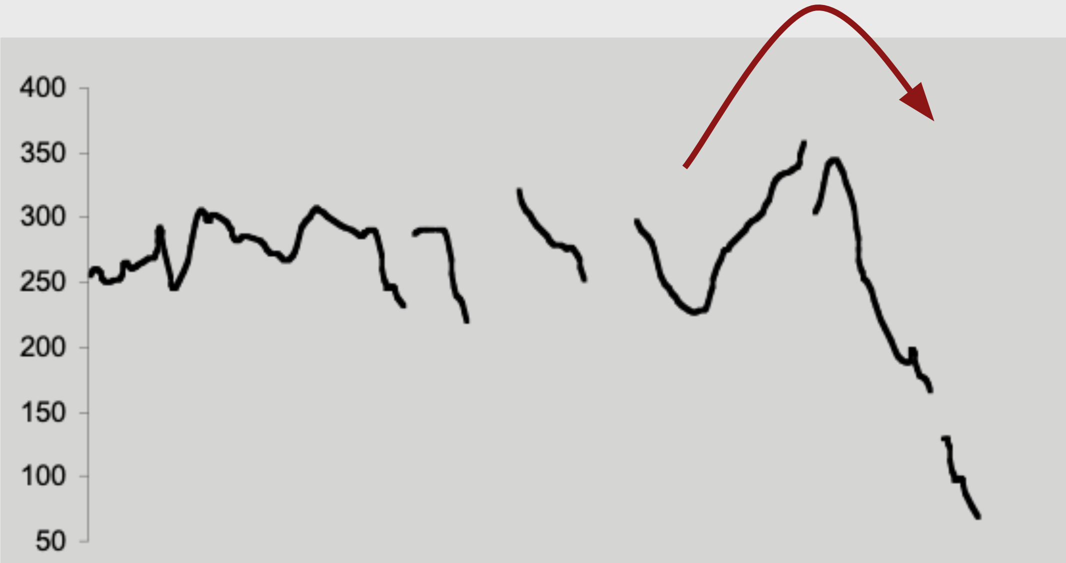
Same ‘Tune’, Different Alignment

- The main **rise-fall** accent (= “I assert this”) shifts locations



Same ‘Tune’, Different Alignment

- The main **rise-fall** accent (= “I assert this”) shifts locations



Legumes are a good source of **VITAMINS**

Predicting Boundaries: Full || versus intermediate |

Ostendorf and Veilleux. 1994 “Hierarchical Stochastic model for Automatic Prediction of Prosodic Boundary Location”, Computational Linguistics 20:1

Computer phone calls, || which do everything | from selling magazine subscriptions || to reminding people about meetings || have become the telephone equivalent | of junk mail. ||

Doctor Norman Rosenblatt, || dean of the college | of criminal justice at Northeastern University, || agrees.||

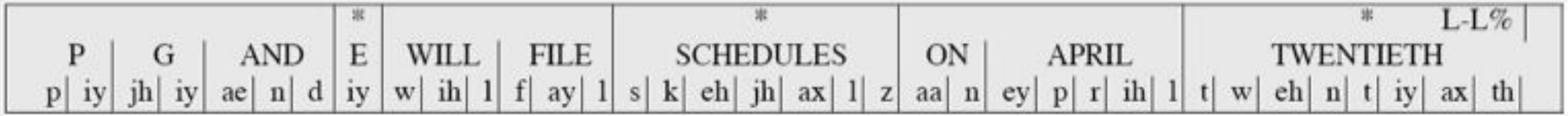
For WBUR, || I’m Margo Melnicove.

Overview of the Speech Synthesis Task

Synthesis in Two Stages

```
PG&E will file schedules on April 20th
```

1. **Text Analysis:** Text into some intermediate representation. *What to say*



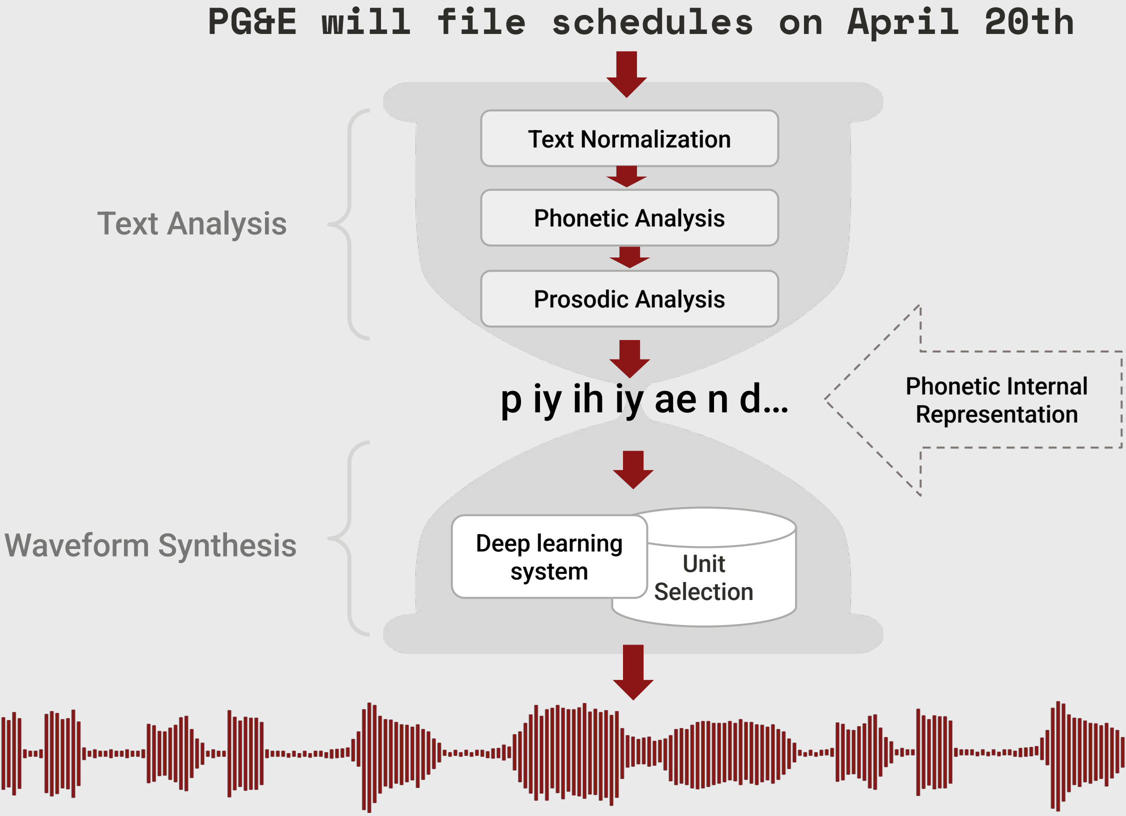
			*			*			* L-L%
P	G	AND	E	WILL	FILE	SCHEDULES	ON	APRIL	TWENTIETH
p iy	jh iy	ae n d	iy	w ih l	f ay l	s k eh jh ax l z	aa n	ey p r ih l	t w eh n t iy ax th

2. **Waveform Synthesis:** From the intermediate representation into audio waveform. *Saying it*

Synthesis in Two Stages

- **Convert messy text, prosodic commands, or other metadata into normalized text**
    - Synthesis tends to sound more consistent when trained on normalized text data. Even with vast training sets.
    - Public models available. LLMs should be great at this! You can use few-shot LLM prompting to normalize.
    - Some design decisions about how to pronounce some written forms. And how to enable prosody controls
- **Key interface: Intermediate representation output by text conversion. Speech audio generates audio from this representation**
- **Waveform generation. Create audio from text + prosodic information**
    - Modern systems: One or more deep learning modules.
    - Older ideas: Unit selection. Stitch together curated examples of actual speech to form new utterances
    - Text -> Mel spectrogram + voicing -> waveform is a somewhat standard set of steps
    - End-to-end is possible – from text directly to waveform (or at least spectrogram + separate vocoder)

Two-stage Approach Allows Separation of Data Collection



PG&E will file schedules on April 20th

Text Analysis

Text Normalization

Phonetic Analysis

Prosodic Analysis

p iy ih iy ae n d...

Phonetic Internal Representation

Waveform Synthesis

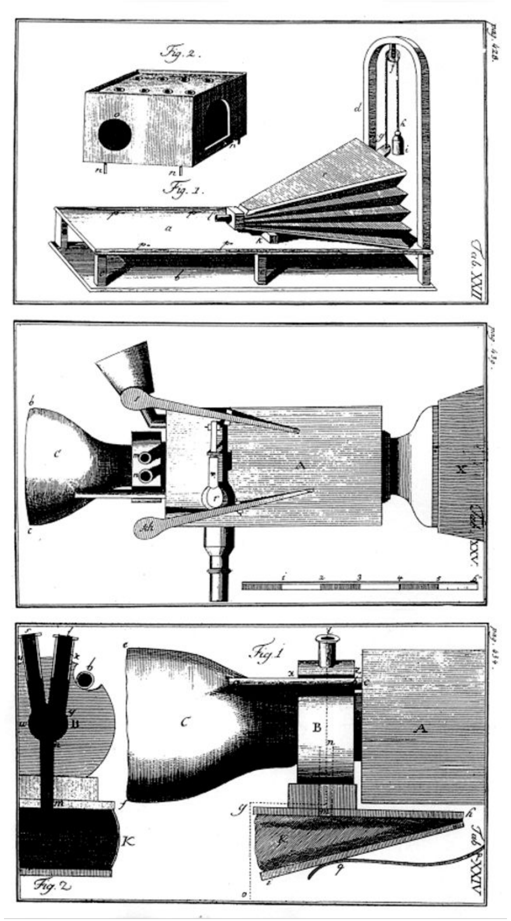
Deep learning system

Unit Selection

TTS Modeling History and Overview

Synthesis in 1780
Von Kempelen

- Small whistles controlled consonants
- Rubber mouth and nose; nose had to be covered with two fingers for non-nasals
- Unvoiced sounds: mouth covered, auxiliary bellows driven by string provides puff of air



From Traunmüller's website

Homer Dudley
1939 VODER

- Manually controlled through complex keyboard
- Operator training was a problem
- A demo trick: Telling the audience what they are about to hear improves understanding

Gunnar Fant’s OVE Synthesizer

- Of the Royal Institute of Technology, Stockholm
- Operator training was a problem, complex controls to play it

Pre-modern TTS

- 1960's first full TTS: Umeda et al (1968)
- Joe Olive 1977 concatenation of linear-prediction diphones
- Texas Instruments Speak and Spell
  - June 1978
  - Paul Breedlove

1990s - ~2015: Improving Unit Selection Synthesis

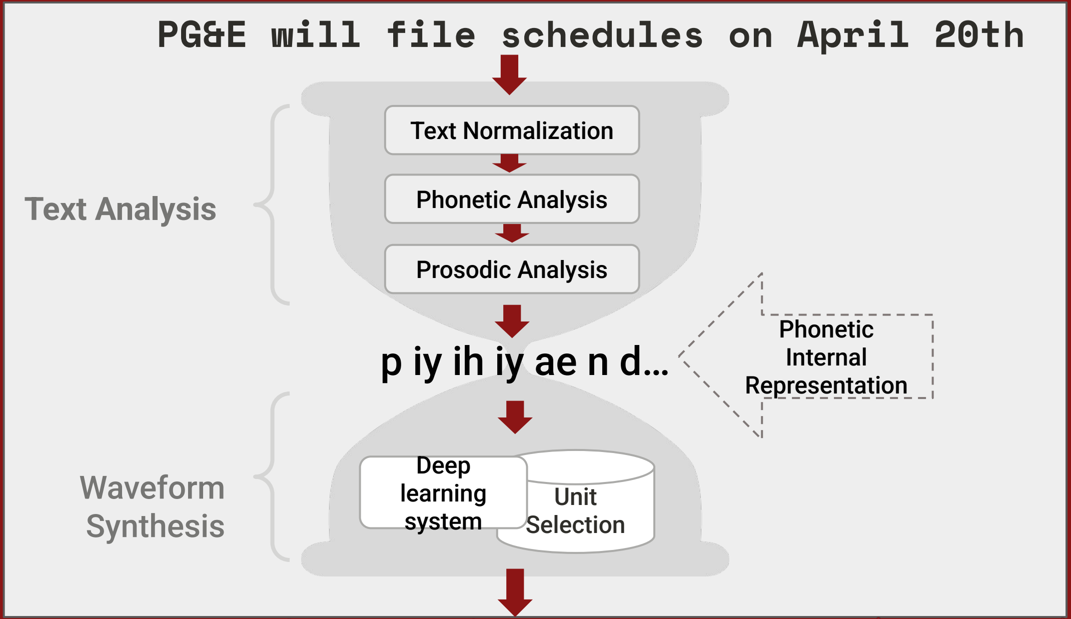
- **Units are diphones or larger word chunks**
    - As computes improved, number of units went from thousands to millions (including specific phrases)
- **Paste them together and modify prosody.**
    - No need to simulate low-level aspects of voice, just use recordings!
    - Generating realistic prosody is hard
    - Systems built on single speaker. No clear way to train on many speakers and generate interpolated or novel voices
- **Small errors occur where units join. Some special heuristics to help smooth joins ("join errors")**
- **Deep learning systems for parametric synthesis improved in early 2010's. Then deep learning took over**
    - Neural nets for TTS go back to at least the 1970's. Dan Jurafsky did a postdoc on this in the early 90's!
    - Parametric synthesis simply didn't have good enough function approximation models until 2010's

Parametric Synthesis

- **Train a statistical model on large amounts of data.**
- **Learn text -> sound mapping.**
  - Possibly use phonemes as an intermediate representation.
  - Deep learning architectures have recently explored different ways to split up this task
- **Previously associated with HMM Synthesis**
  - A reverse of HMM-based speech recognition. Lots of hand-coded assumptions.
- **Deep learning approaches (Wavenet, Tacotron) made parametric a clear winner**
  - Early success of modern deep learning approaches relied on careful audio engineering
  - Use neural net to predict parameters of a hand-engineered vocoder (parameters -> waveform)

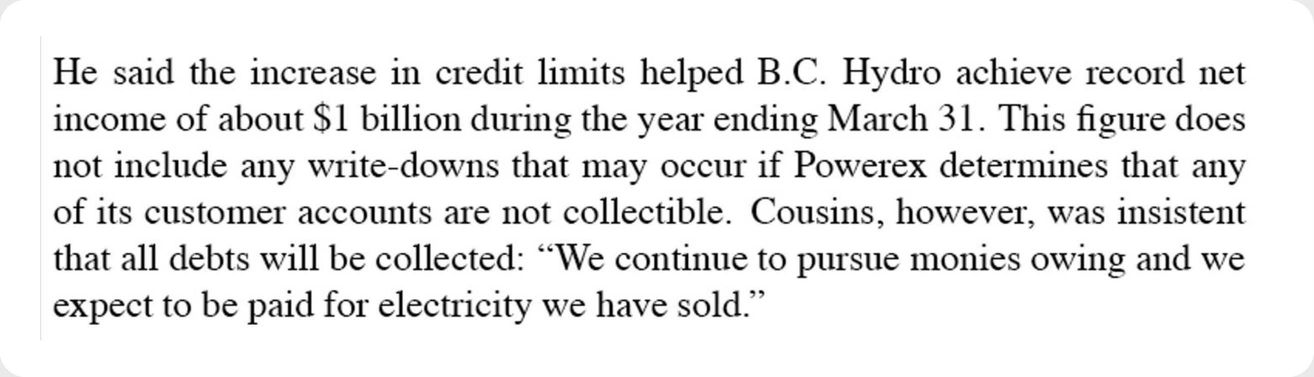
Current trend: Audio-LLM multimodal models, and general promptable audio generation models

Text Normalization



Text Normalization

- Analysis of raw text into pronounceable words:



He said the increase in credit limits helped B.C. Hydro achieve record net income of about $1 billion during the year ending March 31. This figure does not include any write-downs that may occur if Powerex determines that any of its customer accounts are not collectible. Cousins, however, was insistent that all debts will be collected: “We continue to pursue monies owing and we expect to be paid for electricity we have sold.”

- Sentence Tokenization
- Text Normalization
- Identify tokens in text
  - Chunk tokens into reasonably sized sections
  - Map tokens to words
  - Tag the words

Understanding the problem: What should we say out loud?

- He stole $100 million from the bank
- It’s 13 St. Andrews St.
- The home page is http://www.stanford.edu
- Yes, see you the following tues, that’s 11/12/01
- IV: four, fourth, I.V.
- IRA: I.R.A. or Ira
- 1750: seventeen fifty (date, address) or one thousand seven… (dollars)

Text Normalization high level steps

1. **Identify tokens in text (Possibly constrained to pre-trained tokenizers in LLM era)**
    a. Chunk tokens into groups (depending on tokenization choice)

2. **Convert tokens to spoken word forms**
    a. Identify types of tokens and use per-type converters
    b. OR use LLM-style approach and fine tuning data to cover all types relevant to the system

VERY likely you can leverage pre-trained LLMs pre-existing normalization systems to help. This is a recent change so look for newer/updated tools.

Text normalization with modern NLProc models

- Transformers and pre-trained LLMs are great tools for text normalization!
- Requires consistent annotations/schema and edge case handling
- Opportunity + Risk: LLMs can help paraphrase / expand text,
  e.g. “partyInvite20014-2-final.pdf” - > “Party Invite PDF”

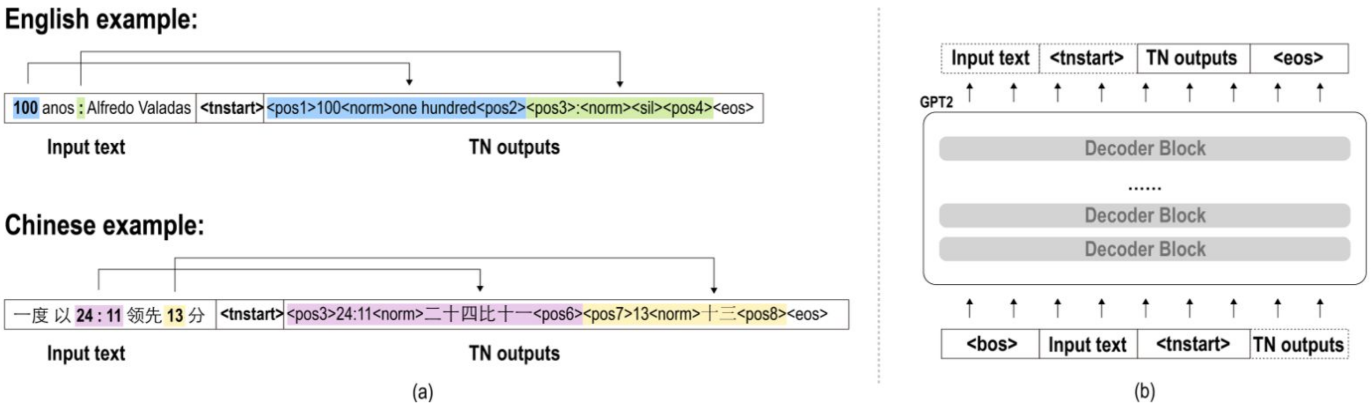


Fig. 1. Architecture of the TNFormer model. (a) two examples of training text sequences, with an English text sequence at the top and a Chinese text sequence at the bottom. (b) the model’s input and output sequences are depicted, with the dashed-line box indicating elements used only during the training phase.

(Shen *et al*, 2024)

Important Issue in Tokenization: End-of-utterance Detection

- Relatively simple if utterance ends in ?!
- But what about ambiguity of "."?
- Ambiguous between end-of-utterance, end-of-abbreviation, and both
  - **My place on Main St. is around the corner.**
  - **I live at 123 Main St.**
  - (Not "**I live at 151 Main St**.")

Identify Types of Tokens, Convert Tokens to Spoken Words

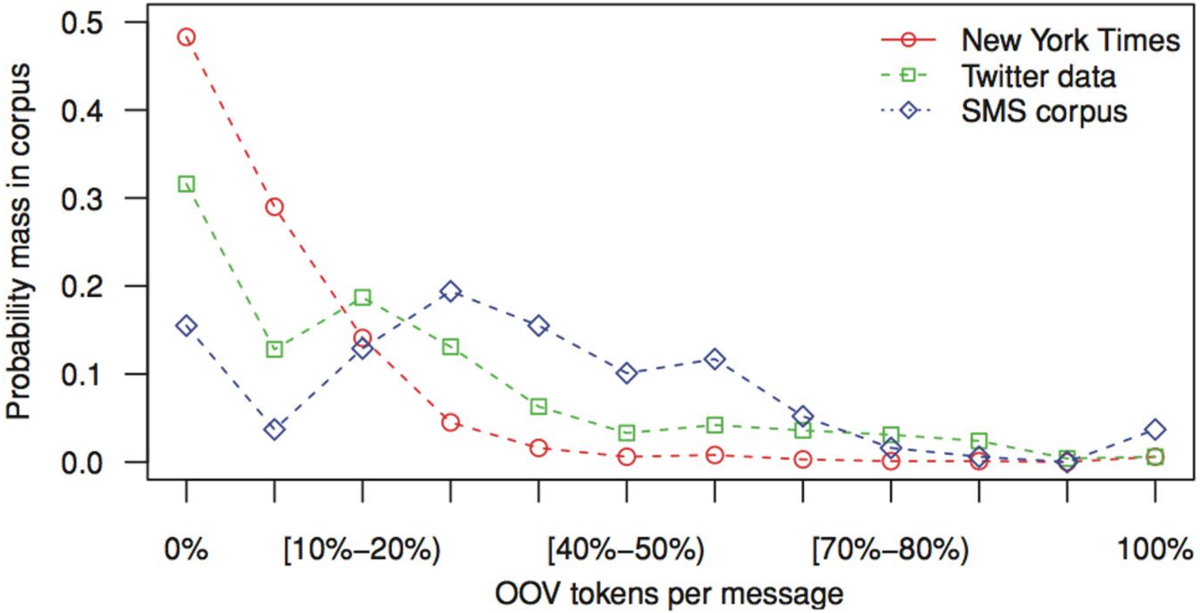
Context matters for accurate disambiguation

- Pronunciation of numbers often depends on type.
- Three ways to pronounce 1776:
  - **Date:** seventeen seventy six
  - **Phone number:** one seven seven six
  - **Quantifier:** one thousand seven hundred (and) seventy six
- Also:
  - 25 Day: twenty-fifth

Can we ignore non-standard words (NSWs)? Frequency analysis

- Word not in gnu aspell dictionary not counting @mentions, #hashtags, urls

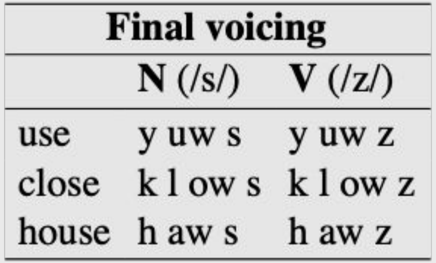
  (Han, Cook, Baldwin 2013)



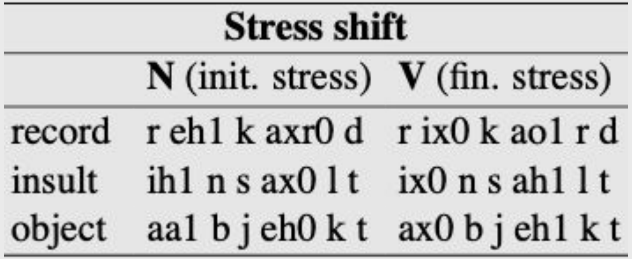
- Twitter: 15% of tweets have 50% or more OOV

Homograph disambiguation

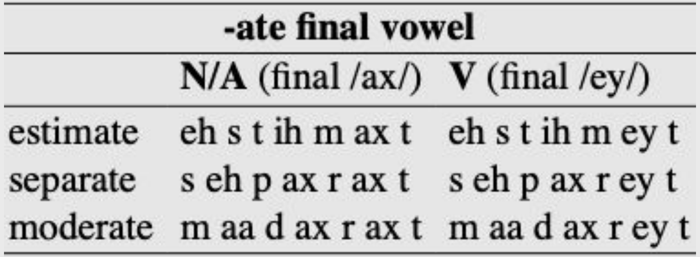
- It's no use (/y uw s/) to ask to use (/y uw z/) the telephone.
- Do you live (/l ih v/) near a zoo with live (/l ay v/) animals?
- I prefer bass (/b ae s/) fishing to playing the bass (/b ey s/) guitar.



Final voicing		
	N (/s/)	**V** (/z/)
use	y uw s	y uw z
close	k l ow s	k l ow z
house	h aw s	h aw z



Stress shift		
	N (init. stress)	**V** (fin. stress)
record	r eh1 k axr0 d	r ix0 k ao1 r d
insult	ih1 n s ax0 l t	ix0 n s ah1 l t
object	aa1 b j eh0 k t	ax0 b j eh1 k t

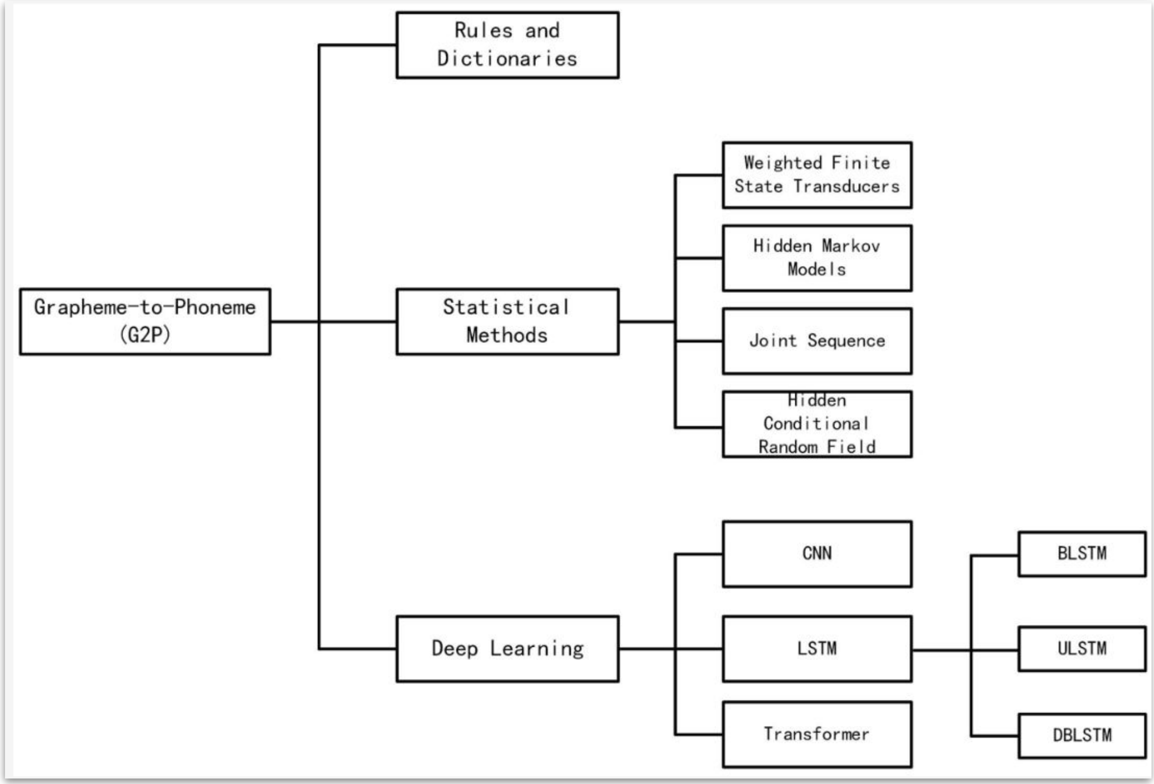


-ate final vowel		
	N/A (final /ax/)	**V** (final /ey/)
estimate	eh s t ih m ax t	eh s t ih m ey t
separate	s eh p ax r ax t	s eh p ax r ey t
moderate	m aa d ax r ax t	m aa d ax r ey t

Mapping from letters to sounds

- AKA Grapheme to Phoneme (G2P)
- Generally machine learning, induced from a dictionary
- How to build: Pick your favorite machine learning tool and go for it
  - There are increasingly many public models trained on large datasets which offer a great default choice
- Modern deep learning methods trained on large datasets remain state-of-the-art
  - Earlier work: (Black et al. 1998) Two steps: alignment and (CART-based) rule-induction
  - Modern DL approaches might handle tokenization and G2P jointly as a single text normalization module
- Choose your G2P system in the context of the overall TTS architecture you're using
  - General tip: You probably don't need to train your own G2P when building a baseline system

Possible types of G2P systems

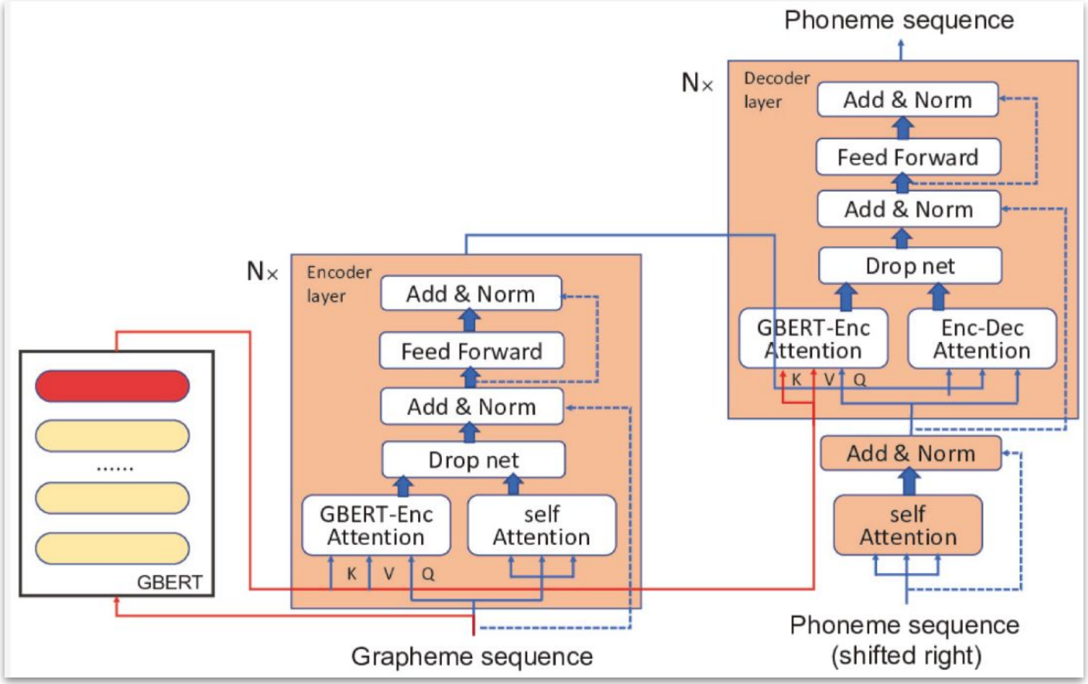


(Cheng, Zhu, Liu, & Wang, 2024)

Cutting edge G2P approaches deep learning and pre-training:

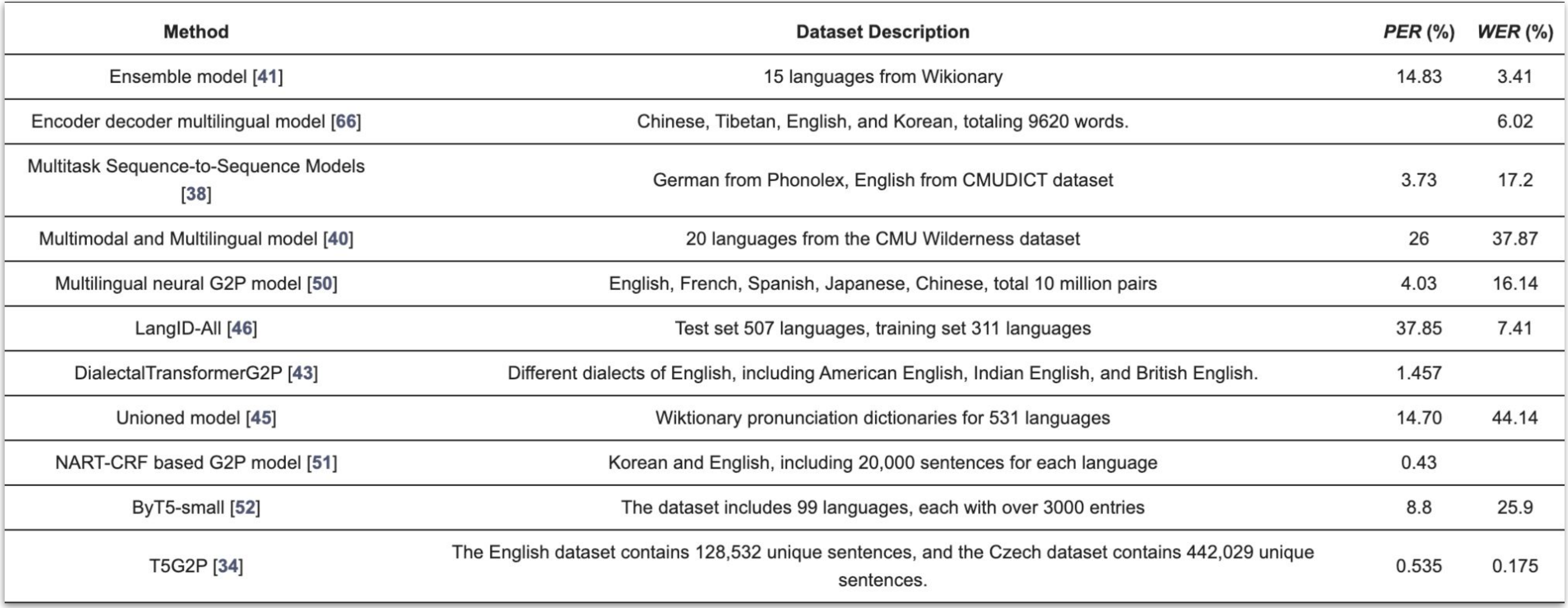
Pre-trained text models + G2P joint data to train specialized modules.

Text data is graphemes, text models can pre-train on vast datasets.



(Cheng, Zhu, Liu, & Wang, 2024)

Evaluation of modern G2P systems on multilingual datasets



Method	Dataset Description	*PER* (%)	*WER* (%)
Ensemble model [41]	15 languages from Wikionary	14.83	3.41
Encoder decoder multilingual model [66]	Chinese, Tibetan, English, and Korean, totaling 9620 words.		6.02
Multitask Sequence-to-Sequence Models [38]	German from Phonolex, English from CMUDICT dataset	3.73	17.2
Multimodal and Multilingual model [40]	20 languages from the CMU Wilderness dataset	26	37.87
Multilingual neural G2P model [50]	English, French, Spanish, Japanese, Chinese, total 10 million pairs	4.03	16.14
LangID-All [46]	Test set 507 languages, training set 311 languages	37.85	7.41
DialectalTransformerG2P [43]	Different dialects of English, including American English, Indian English, and British English.	1.457	
Unioned model [45]	Wiktionary pronunciation dictionaries for 531 languages	14.70	44.14
NART-CRF based G2P model [51]	Korean and English, including 20,000 sentences for each language	0.43	
ByT5-small [52]	The dataset includes 99 languages, each with over 3000 entries	8.8	25.9
T5G2P [34]	The English dataset contains 128,532 unique sentences, and the Czech dataset contains 442,029 unique sentences.	0.535	0.175

(Cheng, Zhu, Liu, & Wang, 2024)

Questions?

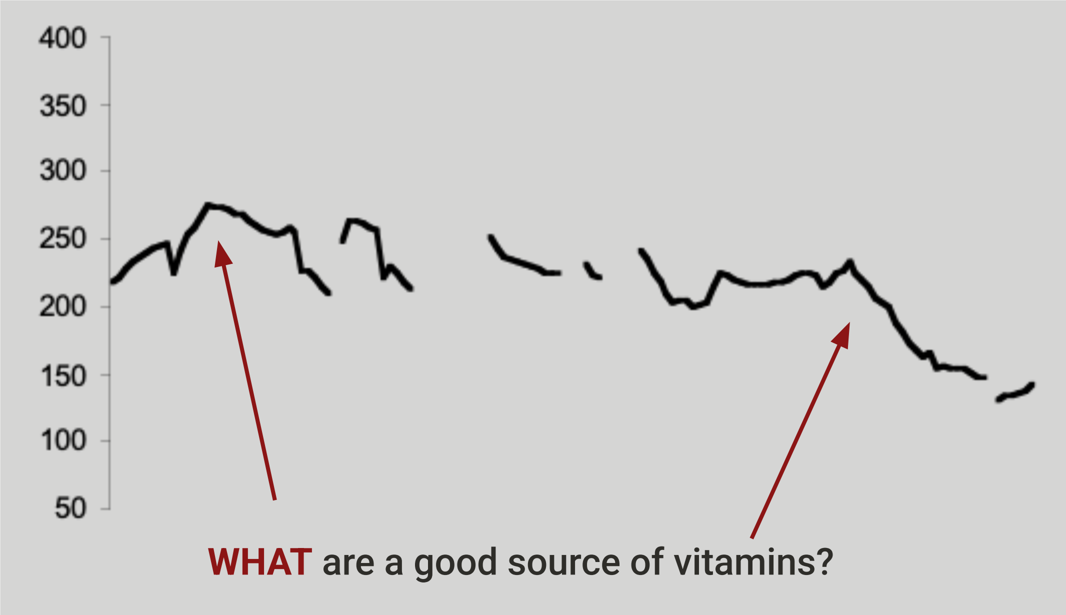
Assignment 1 due in 1 week

More Prosodic Tunes

WH-question Tune

- WH-questions typically have **falling** contours, like statements

[I know that many natural foods are healthy, but ...]

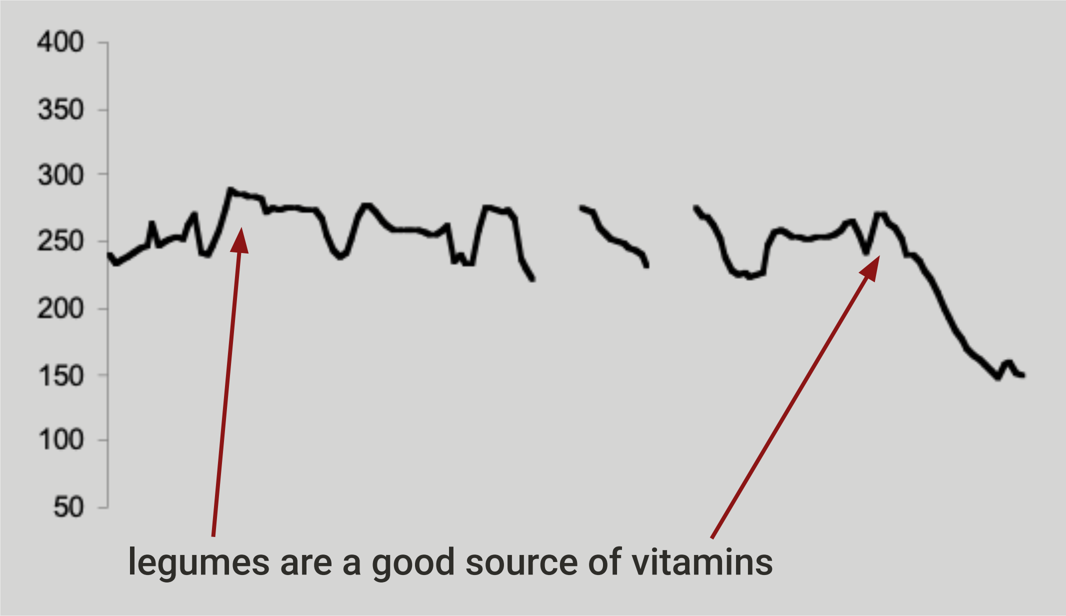


WHAT are a good source of vitamins?

Broad Focus

- In the absence of focus, English tends to mark the **first** and **last** ‘content’ words with prominent accents

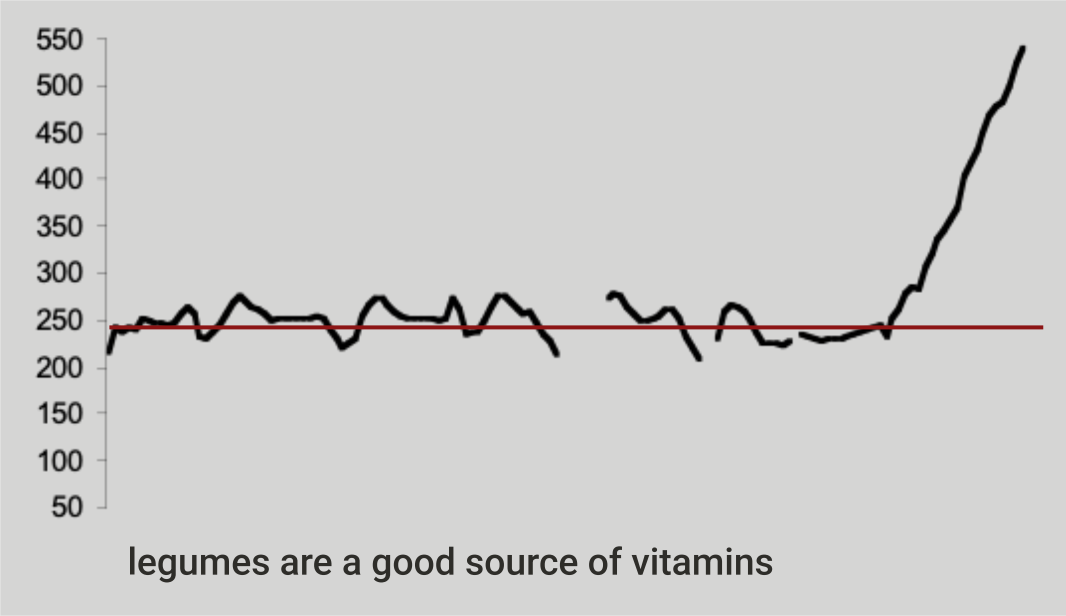
“Tell me something about the world”



Rising Statements

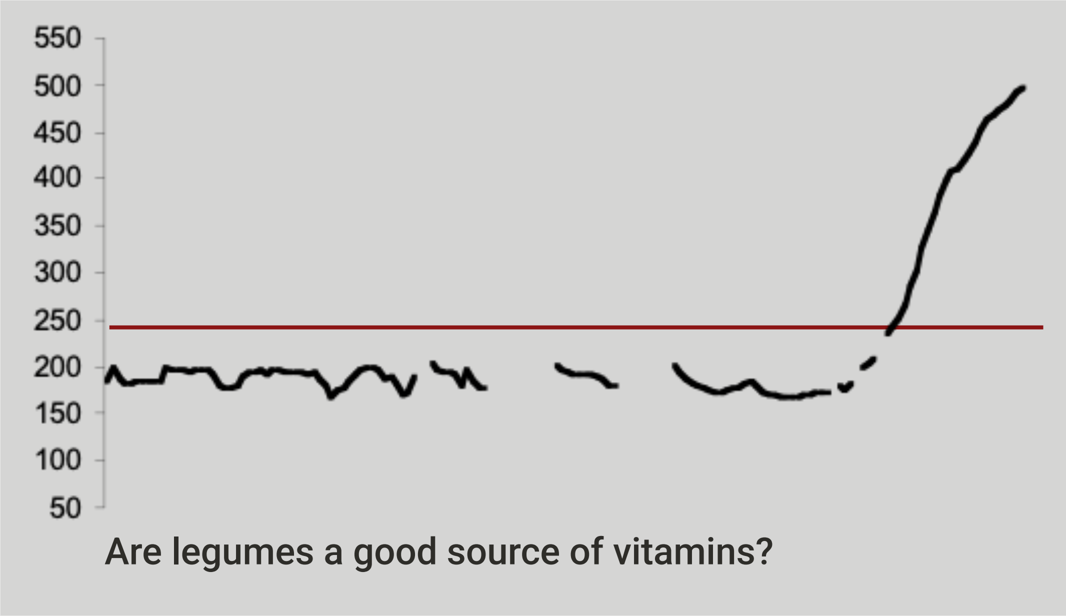
- **High-rising** statements can signal that the speaker is seeking approval

“Tell me something I didn’t know”



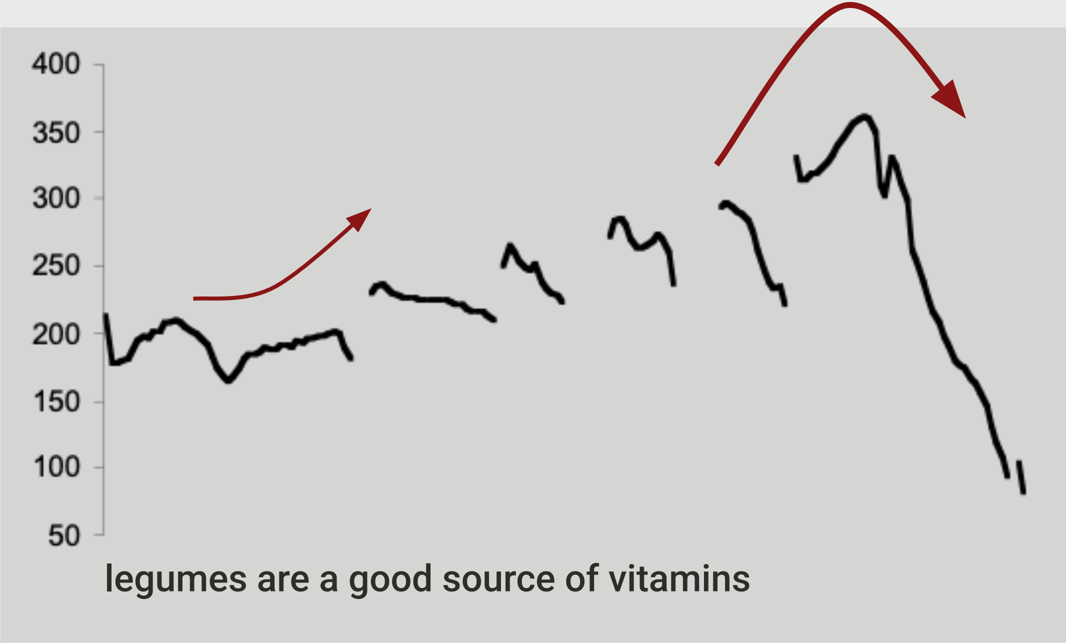
Yes-No question

- **Rise** from the main accent to the end of the sentence



‘Surprise-redundancy’ Tune

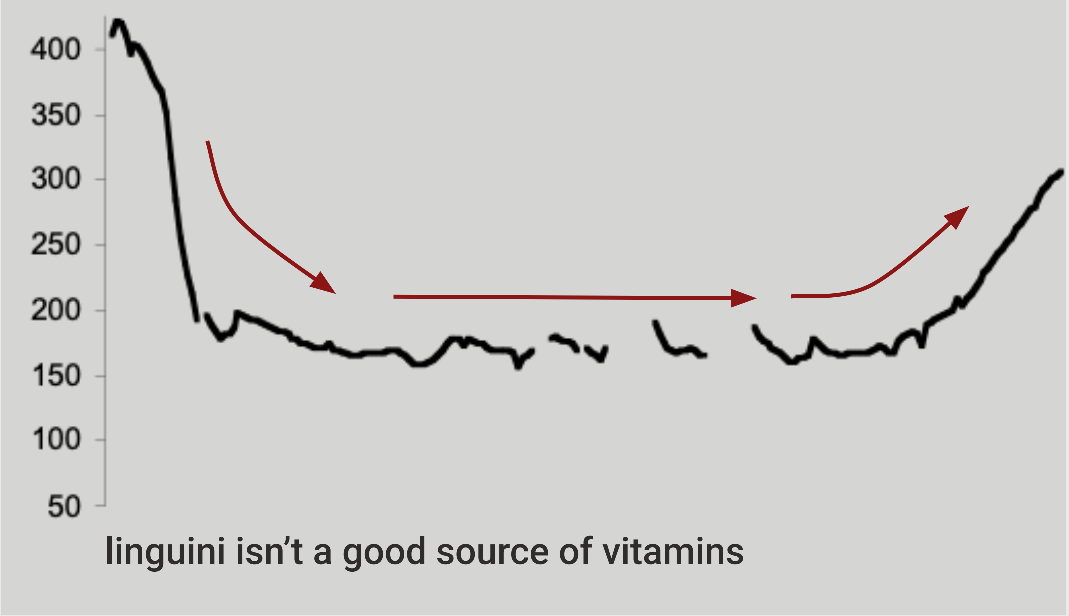
- **Low** beginning followed by a gradual rise to a **high** at the end



‘Contradiction’ Tune

- **Sharp fall** at the beginning, **flat and low**, then **rising** at the end

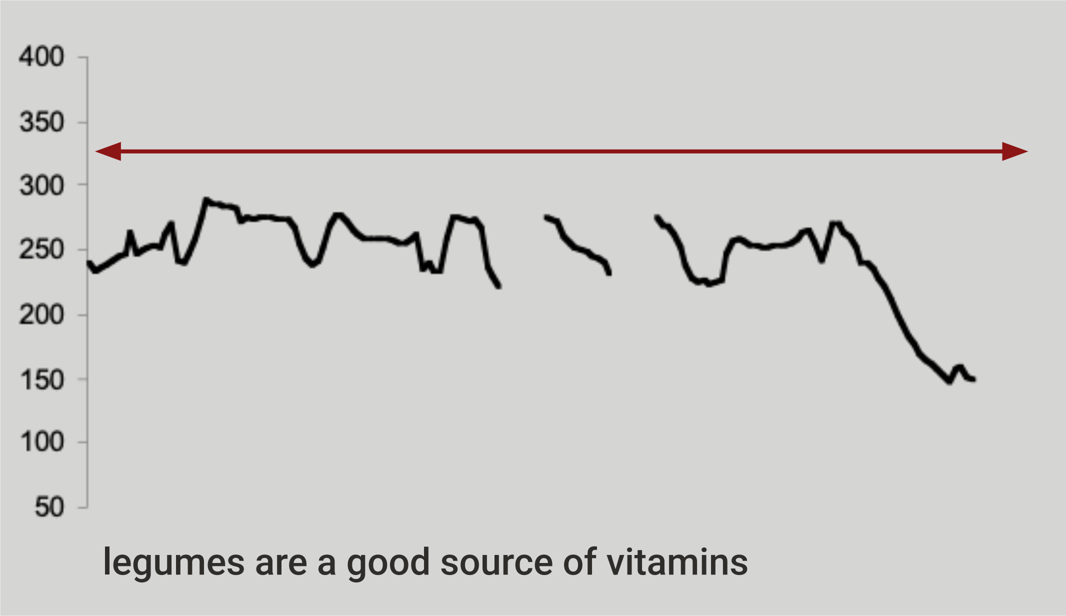
“I’ve heard that linguini is a good source of vitamins”



[... how could you think that?]

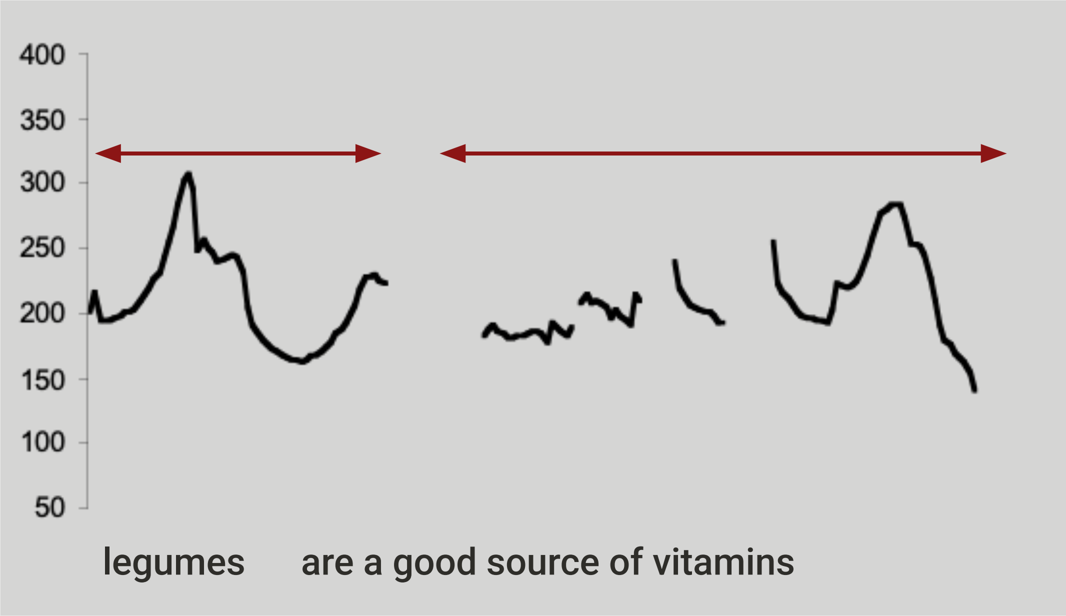
A Single Intonation Phrase

- Broad focus statement consisting of one intonation phrase (that is, one intonation tune spans the whole unit).



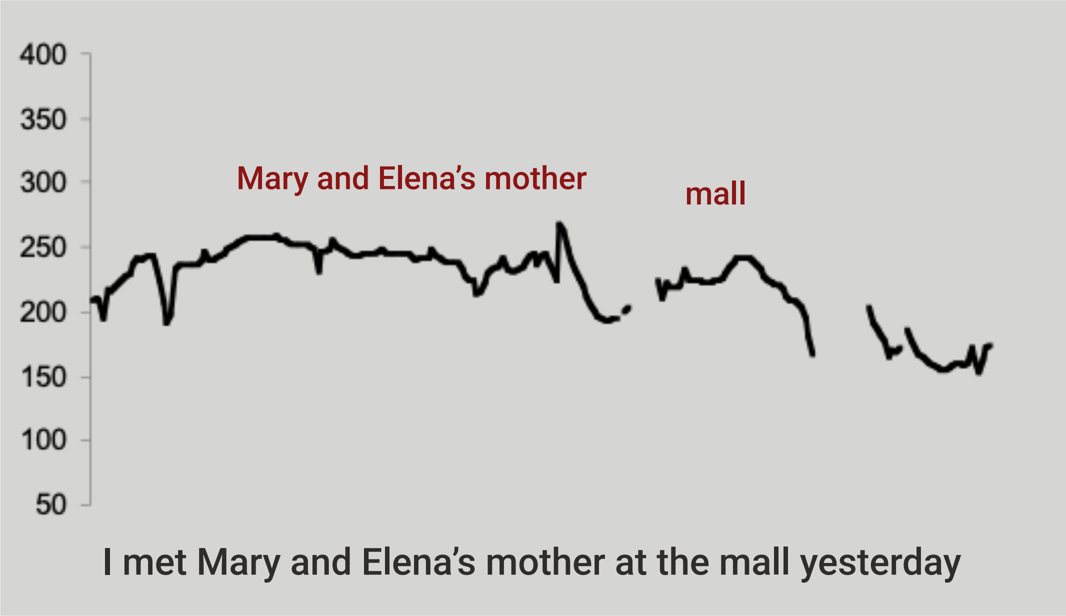
Multiple Phrases

- Utterances can be ‘chunked’ up into smaller phrases in order to signal the importance of information in each unit.



Phrasing Sometimes Helps Disambiguate

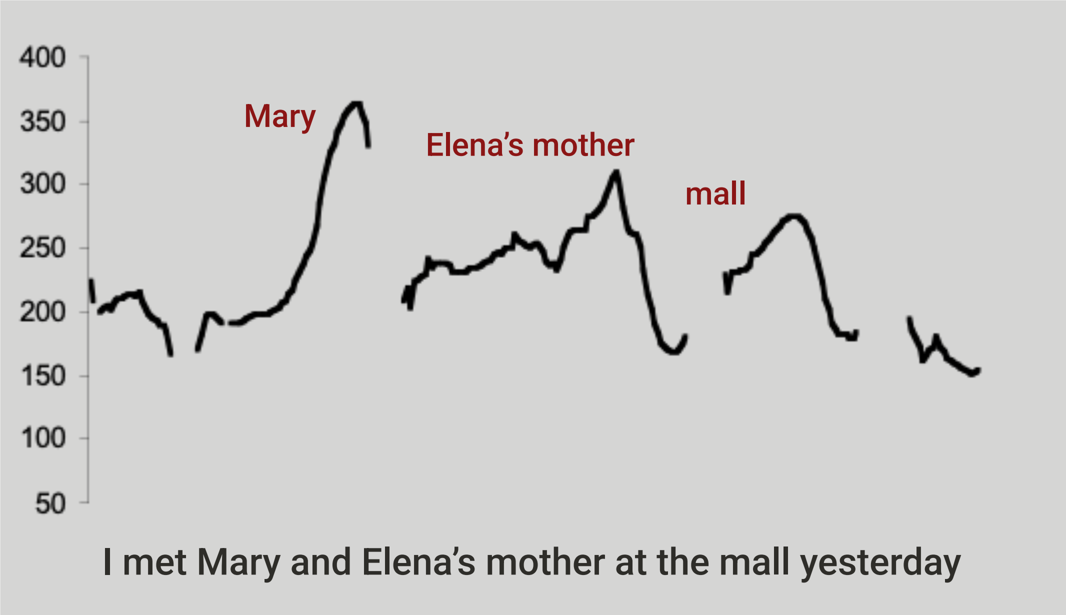
- One intonation phrase with relatively flat overall pitch range



I met Mary and Elena's mother at the mall yesterday

Phrasing Sometimes Helps Disambiguate

- Separate phrases, with expanded pitch movements



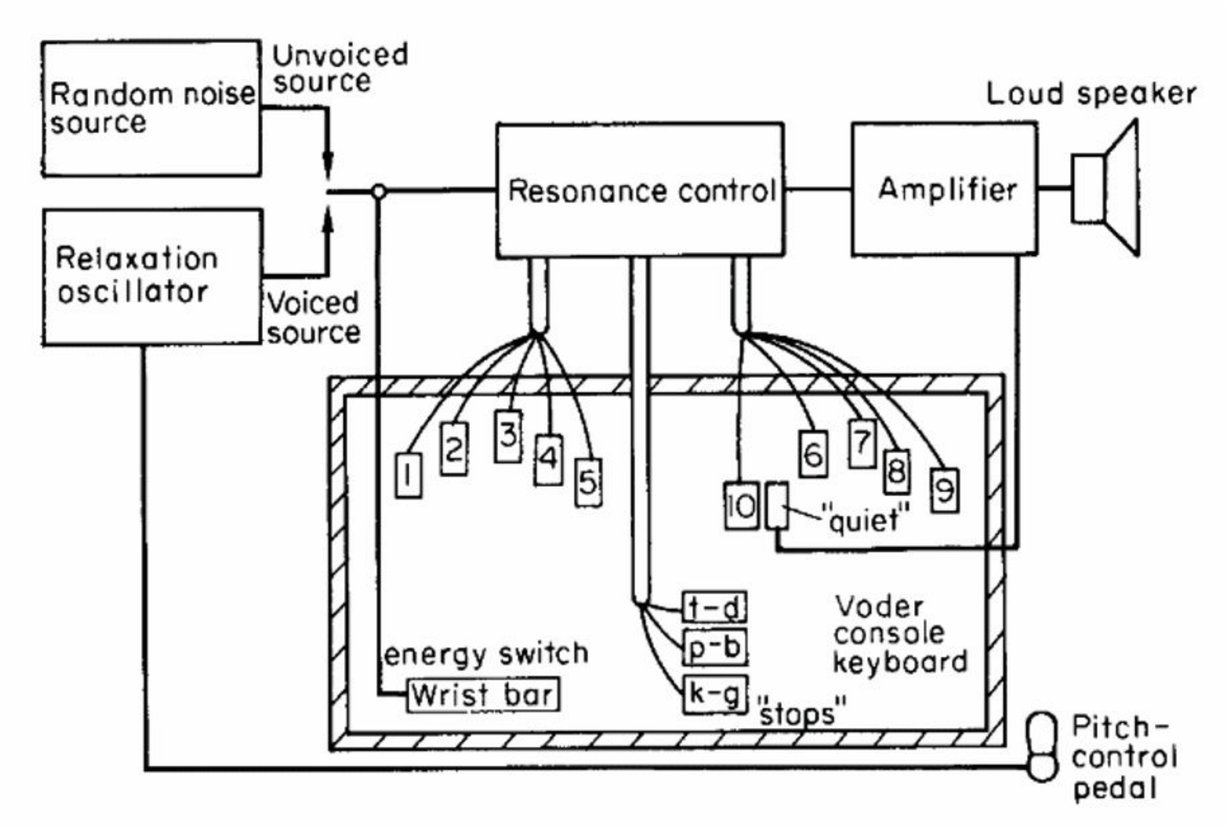
Levels of Prominence

- Most phrases have more than one accent
- **Nuclear Accent:** Last accent in a phrase, perceived as more prominent
  - Plays semantic role like indicating a word is contrastive or focus.
  - Modeled via ***s in text messages, or all-capitalized letters
  - ‘I know **SOMETHING** interesting is sure to happen,’ she said
- Can also have reduced words that are less prominent than usual (especially function words)
- Design choice: How to model prominence. Sometimes use 4 classes:
  - Emphatic accent, pitch accent, unaccented, reduced

Appendix

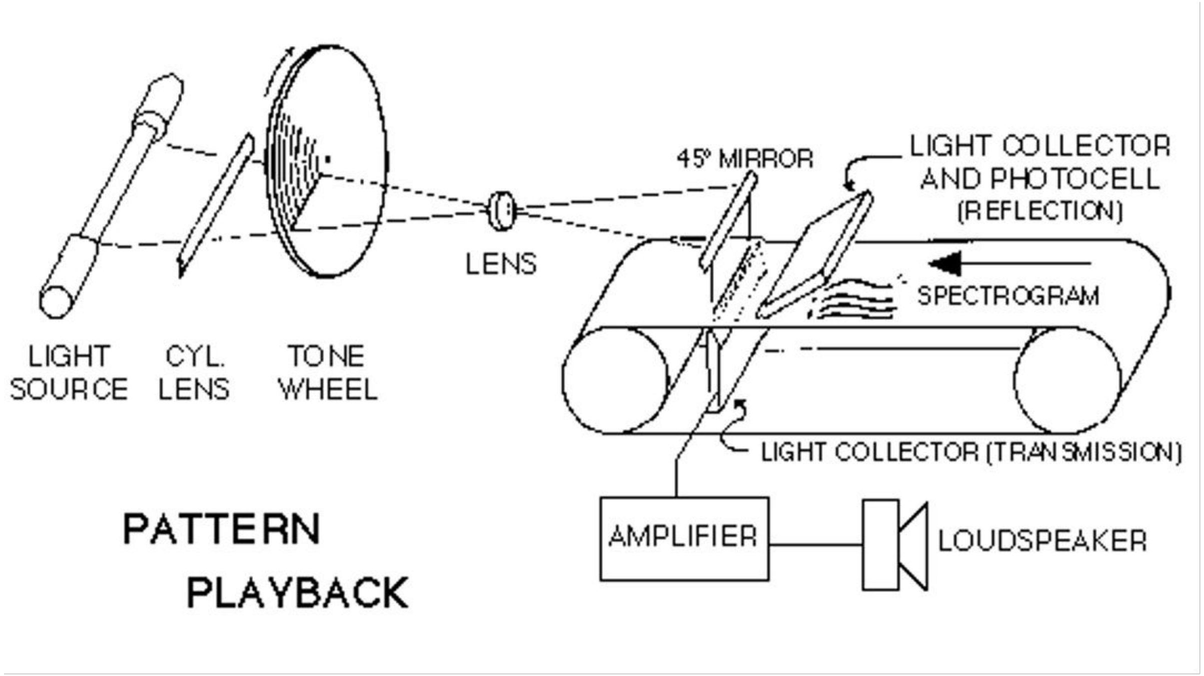
Homer Dudley
1939 VODER

- Synthesizing speech by electrical means
- 1939 World's Fair



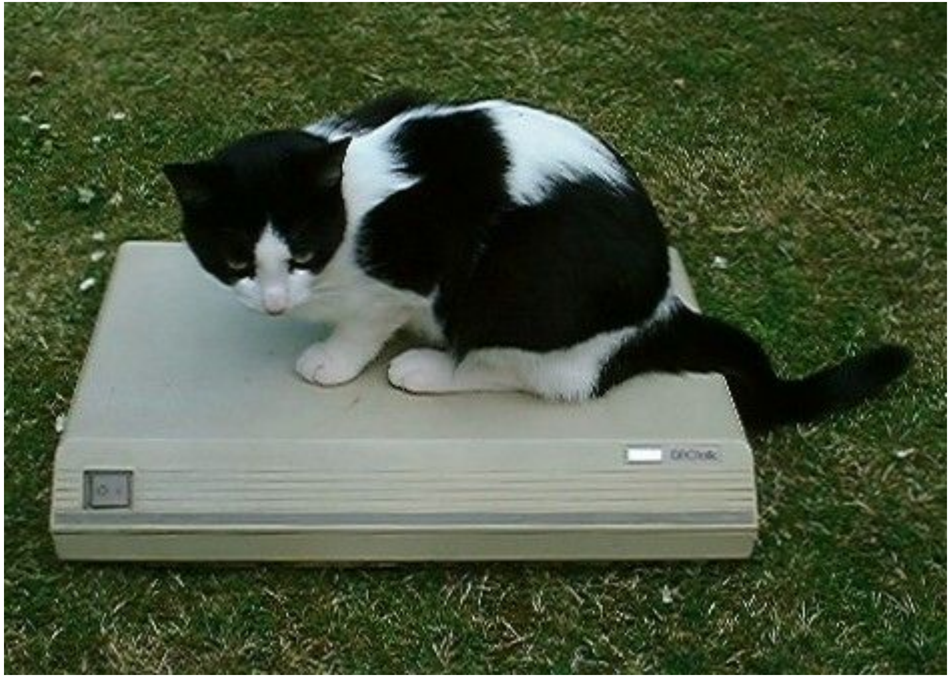
Cooper's Pattern Playback

- Haskins Labs for investigating speech perception
- Works like an inverse of a spectrograph
- Light from a lamp goes through a rotating disk then through spectrogram into photovoltaic cells
- Thus amount of light that gets transmitted at each frequency band corresponds to amount of acoustic energy



1980s: Formant Synthesis

- Were the most common commercial systems when computers were slow and had little memory.
- 1979 MIT MITalk (Allen, Hunnicut, Klatt)
- 1983 DECtalk system based on Klatttalk
    - "Perfect Paul" (The voice of Stephen Hawking)
    - "Beautiful Betty"



Festival

- Open source speech synthesis system
- Designed for development and runtime use
  - Use in many commercial and academic systems
  - Hundreds of thousands of users
- Multilingual
  - No built-in language
  - Designed to allow addition of new languages
- Additional tools for rapid voice development
  - Statistical learning tools
  - Scripts for building models

Slide: Richard Sproat

Festival as Software

- http://festvox.org/festival/
- General system for multilingual TTS
- C/C++ code with Scheme scripting language
- General replaceable modules:
  - Lexicons, LTS, duration, intonation, phrasing, POS tagging, tokenizing, diphone/unit selection, signal processing
- General tools
  - Intonation analysis (f0, Tilt), signal processing, CART building, N-gram, SCFG, WFST

Slide: Richard Sproat

Identify Tokens and Chunk in older systems

- Whitespace can be viewed as separators
- Punctuation can be separated from the raw tokens
- For example, Festival (older research TTS system) converts text into
    - ordered list of tokens
    - each with features:
        - its own preceding whitespace
        - its own succeeding punctuation

CMU FestVox Project

- **Festival is an engine, how do you make voices?**
- **Festvox: building synthetic voices:**
    - Tools, scripts, documentation
    - Discussion and examples for building voices
    - Example voice databases
    - Step by step walkthroughs of processes
    - Support for English and other languages
- **Support for different waveform synthesis methods**
    - Diphone
    - Unit selection

Slide: Richard Sproat

Dictionaries

- FCMU dictionary: 127K words
  - http://www.speech.cs.cmu.edu/cgi-bin/cmudict
- Unisyn dictionary
  - Significantly more accurate, includes multiple dialects
  - http://www.cstr.ed.ac.uk/projects/unisyn/

Slide: Richard Sproat

How Common are Non-standard Words (NSWs)?

- Word not in lexicon, or with non-alphabetic characters (Sproat et al 2001, before SMS/Twitter)

Text Type	%NSW
novels	1.5%
press wire	4.9%
e-mail	10.7%
recipes	13.7%
classifieds	17.9%

Names

- Big problem area is names
- Names are common
  - 20% of tokens in typical newswire text
  - Spiegel (2003) estimate of US names:
    - 2 million surnames
    - 100,000 first names
  - Personal names: McArthur, D'Angelo, Jimenez, Rajan, Raghavan, Sondhi, Xu, Hsu, Zhang, Chang, Nguyen
  - Company/Brand names: Infinit, Kmart, Cytyc, Medamicus, Inforte, Aaon, Idexx Labs, Bebe

Homograph Disambiguation

- 19 most frequent homographs, from Liberman and Church 1992
- Counts are per million, from an AP news corpus of 44 million words.
- Not a huge problem, but still important

use	319	survey	91
Increase	230	project	90
close	215	separate	87
record	195	present	80
house	150	read	72
contract	143	subject	68
lead	131	rebel	48
live	130	finance	46
lives	105	estimate	46
protest	94		

Part of Speech Tagging for Homograph Disambiguation

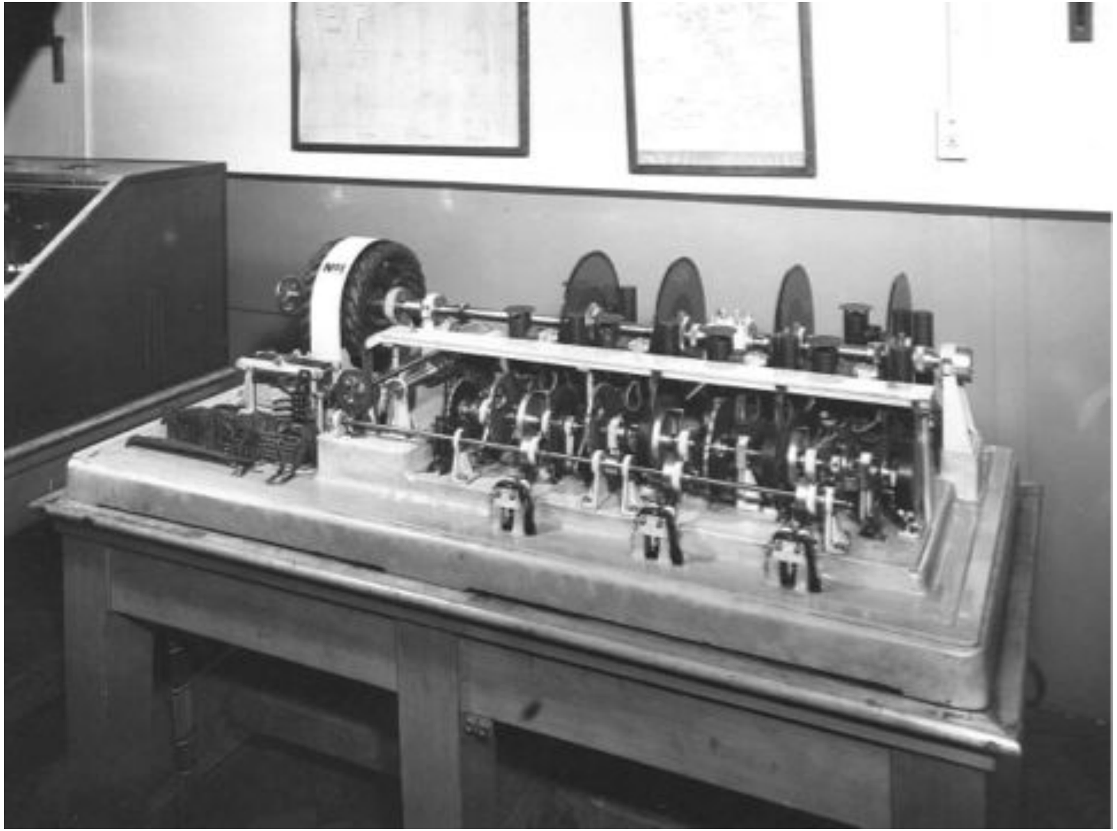
- Many homographs can be distinguished by POS

use	y uw s	y uw z
close	k l ow s	k l ow z
house	h aw s	h aw z
live	l ay v	l ih v
REcord	reCORD	
INsult	inSULT	
OBject	obJECT	
OVERflow	overFLOW	
DIScount	disCOUNT	
CONtent	conTENT	

- POS tagging also useful for CONTENT/FUNCTION distinction, which is useful for phrasing

The 1936 UK Speaking Clock

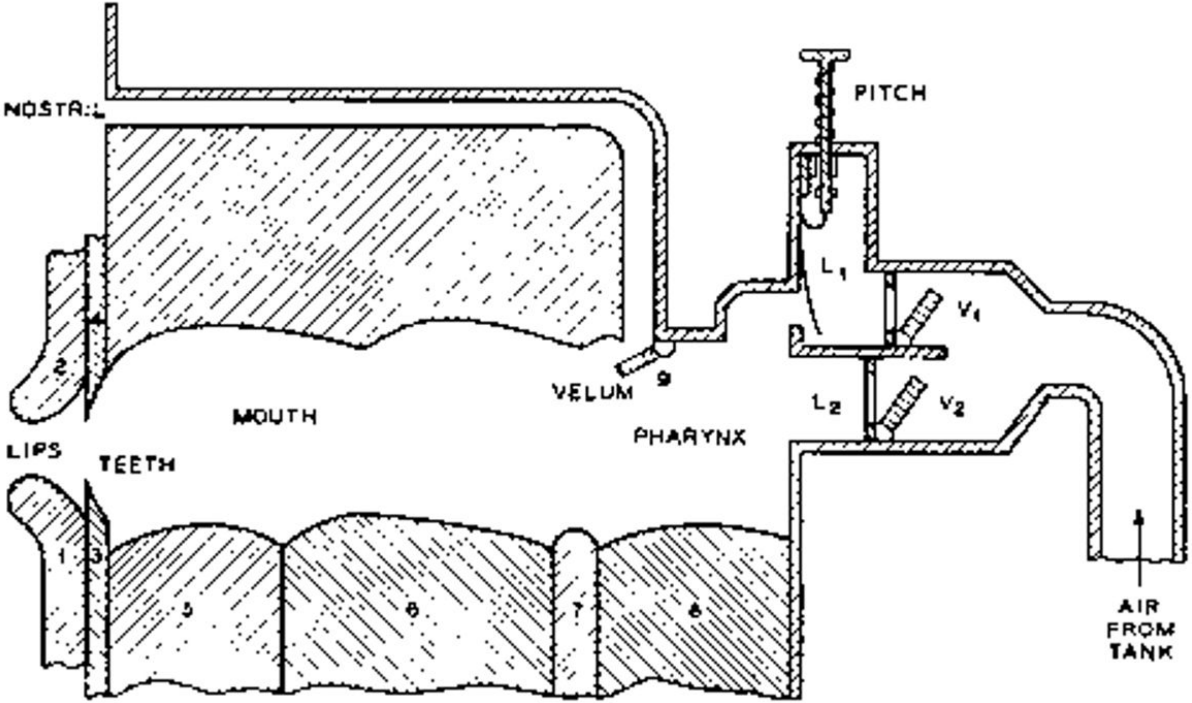
- July 24, 1936
- Photographic storage on 4 glass disks
- 2 disks for minutes, 1 for hour, one for seconds.
- Other words in sentence distributed across 4 disks, so all 4 used at once.
- Voice of “Miss J. Cain”



The 1936 UK Speaking Clock

- July 24, 1936
- A technician adjusts the amplifiers of the first speaking clock

Closer to a Natural Vocal Tract: Riesz 1937


NOSTRIL
PITCH
L1
V1
2
VELUM
9
L2
V2
MOUTH
PHARYNX
LIPS
TEETH
1
3
5
6
7
8
AIR FROM TANK